Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition

Pascal Denis, Philippe Muller

Predicting Globally-Coherent Temporal Structures from Texts via Endpoint Inference and Graph Decomposition

Pascal Denis[†]
† Alpage, INRIA & Univ. Paris Diderot
pascal.denis@inria.fr

Philippe Muller[†,◇]
◇ IRIT, Univ. Toulouse
muller@irit.fr

Abstract

An elegant approach to learning temporal orderings from texts is to formulate this problem as a constraint optimization problem, which can be then given an exact solution using Integer Linear Programming. This works well for cases where the number of possible relations between temporal entities is restricted to the mere precedence relation [Bramsen *et al.*, 2006; Chambers and Jurafsky, 2008], but becomes impractical when considering all possible interval relations. This paper proposes two innovations, inspired from work on temporal reasoning, that control this combinatorial blow-up, therefore rendering an exact ILP inference viable in the general case. First, we translate our network of constraints from temporal intervals to their endpoints, to handle a drastically smaller set of constraints, while preserving the same temporal information. Second, we show that additional efficiency is gained by enforcing coherence on particular subsets of the entire temporal graphs. We evaluate these innovations through various experiments on TimeBank 1.2, and compare our ILP formulations with various baselines and oracle systems.

1 Introduction

Learning the temporal ordering over events, dates and other temporal entities in a text consists in finding a set of temporal relations (precedence, inclusion, etc.) between these entities. This task is an important aspect of discourse understanding and its automation has potential applications (e.g., discourse parsing, text summarization, information extraction).

An important challenge of this task is that temporal relations carry algebraic properties reflecting the linear structure of time (e.g., the transitivity of precedence and inclusion) which make the determination of the temporal relations between entity pairs strongly interdependent. While prior work acknowledges this, most recent approaches to temporal ordering assume a fairly idealized setting, wherein [Mani *et al.*, 2006]:

a. the pairs of entities to be related by the system have been pre-selected by an oracle

b. each of these pairs are labeled independently by a locally-trained classifier

Predicting temporal relations this way (even under assumption (a)) runs the risk of producing structures that are incoherent at the level of the text, which are of little applicative use (especially, if further reasoning is performed on them).

There are a few exceptions to this methodology [Bramsen *et al.*, 2006; Tatu and Srikanth, 2008; Chambers and Jurafsky, 2008]. These approaches directly exploit the inferential properties of the temporal relations to constrain the classifier decisions in a way that ensures overall coherence. Specifically, the following scenario is assumed: (i) learn a soft classifier which outputs a score for each local pair and relation and (ii) combine these local preferences with coherence constraints on the temporal graph within a global optimization problem. Both exact and approximated inference schemes have been investigated for the second step. To perform exact inference, [Bramsen *et al.*, 2006; Chambers and Jurafsky, 2008] propose to use Integer Linear Programming (ILP). This framework has interesting properties: temporal constraints can be encoded in a declarative fashion and efficient solvers are available off-the-shelf. Moreover, ILP has been shown to outperform greedy inference algorithms on this task [Bramsen *et al.*, 2006].

An important restriction used in both papers is to consider only the strict precedence relation (i.e., *before*, *after*). This restriction guarantees formulations that remain manageable by current solvers, but at the expense of expressiveness. Many situations described in texts (e.g., inclusion, overlap) cannot be represented with precedence alone, and it is not clear how this restriction can be mixed with later generalizations.

The standardized TimeML annotation [Pustejovsky *et al.*, 2005] has indeed 12 relations, the interval algebra of [Allen, 1983] has 13, with more complex interactions that can lead to increased combinatorial complexity. Being an NP-hard problem, inference in ILP is highly sensitive to the number of variables and constraints used to represent our problem. In the general case these numbers are both exponential in the number of temporal relations that are used.

In this paper, we explore another strategy that circumvents the complexity problem but in a way that preserves all the information provided by annotations. First, we re-express the network of temporal constraints on intervals into constraints on their endpoints (i.e., from 13 to 3 simple relations), thus dramatically reducing the number of variables and constraints. Crucially, this conversion leads to a much simpler optimiza-

tion problem and the result of the optimization can be "unpacked" without loss of information. Second, we decompose the set of temporal entities in "meaningful" sub-graphs and maintain coherence only within these substructures.

We report on two main sets of experiments on the TimeBank dataset and confirm that the ILP strategy performs well on the task of predicting consistent relations from a local classifier, both when the relevant event pairs are known in advance and when they aren't. Our contribution is thus two-fold: we show how to generalize ILP for temporal prediction with the full set of relations used in available annotations, and we provide the first attempt at solving the task without any of the relaxed constraints assumed in prior work.

The rest of the paper is organized as follows. Section 2 presents the context of this study (data, representations, related work). Section 3 details our translation of the problem into point representations using ILP, and Section 4 develops our methods for decomposing the global temporal problem. Finally we discuss our experiments and results in Section 5.

2 Background

The problem of constructing temporal structures from texts is illustrated on a small, slightly simplified example from the Aquaint TimeML corpus:

> President Joseph Estrada on Tuesday$_{t_4}$ condemned$_{e_1}$ the bombings$_{e_5}$ of the U.S. embassies in Kenya and Tanzania and offered$_{e_{12}}$ condolences to the victims. [...] In all, the bombings$_{e_{10}}$ last week$_{t_5}$ claimed$_{e_4}$ at least 217 lives.

The fully specified temporal graph for this document is given in Figure 1 with the 5 original human annotated links (in light shade) and a relation between dates based on a simple calculation from the annotation (in dark shade); it is augmented with the 6 relations (or 17 if we include inverses) that can be inferred from it.

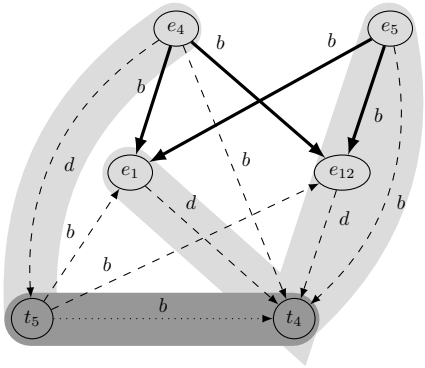

Figure 1: The full temporal graph for the Aquaint excerpt. Labels *b* and *d* stand for *before* and *during*.

The graph exhibits 3 different types of relations: event-event (E-E, solid lines) relations, event-time (E-T, dashed) relations, and time-time (T-T, dotted) relations. In line with previous work, our focus will be on predicting E-E relations, but we will however make use of E-T and T-T information to ensure global coherence (more on this in Section 5).

2.1 Data and representation

Most recent research efforts on temporal processing have been based on TimeBank.[1] This corpus (version 1.2) contains 186 news report documents collected from the DUC and ACE evaluation campaigns, annotated using the TimeML standard for tagging *events* (and states) and time expressions (*timex* for short), including dates and durations, and their temporal relations (called *tlinks*). There are 11 types of relation, whose semantics is similar to the well known Allen relations on intervals [Allen, 1983], excluding the *overlap* relation and its inverse. Allen relations defines all possible ways of relating two intervals according to all possible orderings of their endpoints. The distribution of these relations on E-E pairs is summarized in Table 1, with Allen and TimeML names. This table includes the additional E-E relation instances proposed by [Bethard *et al.*, 2007].

The "Base" column lists the relations annotated in the corpus, while the "Sat." column lists the relations that can be deduced from the annotations with the proper inference procedure (or "saturation") explained below. We also added relations between dates that are implicit because of the values put by the annotator (in the example, t_4 had the value `1998-08-11` and t_5 `1998-08-04`). Combined with other E-T and E-E relations, these generate also new E-E relations. Each relation has an inverse relation (not shown in the table) except *equals*. The majority class in each case (28 and 37%) is thus based on twice as many relations.

Allen	TimeML	Base (28%)	Sat. (37%)
b(efore)	*before*	785	12053
e(quals)	*simul.*, *ident.*	1666	2462
d(uring)	*during*,*incl'ed*	370	1303
f(inish)	*ends*	43	82
s(tart)	*begins*	41	72
m(eet)	*ibefore*	39	78
o(verlap)	n/a	0	1

Table 1: Event-Event relation distribution in TimeBank 1.2, before and after saturation, with % for the majority class.

In order to make explicit the relations that are implicit in human annotated corpora, a procedure is needed to combine the already existing information in the most precise way. Initially, authors have relied on composition rules for simple relations only but turned to more general models dedicated to temporal reasoning [Verhagen, 2005]. The most appropriate in the context of annotation of event relation is the interval algebra of [Allen, 1983], as TimeML relations have the same semantics as a subset of Allen relations. A relation algebra defines a calculus on a set of base relations and any disjunctions of these relations, with union and intersection operators on disjunctions, and a composition of relations. Knowing relations between (x,y) and (y,z), one can compose them to add constraints on the relation bearing on (x,z), only using a table of composition of the 13 base relations, union and intersection. Applying this to all possible triplets of relation of a graph of temporal constraints is known as path-consistency

[1] `http://timeml.org/site/timebank/`

checking, and can enrich a representation as seen Table 1, but can also detect an inconsistent set of constraints, which is crucial in the context of relation prediction. In the general case, path-consistency checking is correct but not complete: some inconsistent configurations cannot be detected that way. Fortunately, in the case of human temporal annotations, only simple relations are initially present, and their compositions generate a sub-algebra of so-called "convex" relations over which path-consistency is complete [Van Beek, 1990].

2.2 Evaluation

The evaluation methodology of studies that consider relations independently is simply to estimate the accuracy of the classifier on the annotated event pairs. To take coherence into account and address the problem at the text level, the evaluation must also consider the whole graph of E-E pairs, including relations that can be inferred from the reference. In the example, it means we must provide an answer for the relation $e_5 < e_1$, which was not annotated but can be inferred from the annotation. The methodology followed in [Mani *et al.*, 2006] is to propagate constraints in the graph and evaluate system and reference with respect to all the event pairs that end up related with a simple TimeML relation in the reference and the system, yielding precision and recall scores. In a global setting, a more complete methodology should also evaluate the consistency of the graphs produced.

2.3 Previous work

Research on temporal ordering has a long history in NLP. Early work was mainly concerned with the study of language mechanisms and information sources (such as tense, aspect, lexical semantics, rhetorical relations) that impact temporal ordering. The recent availability of annotated ressources like the TimeBank corpus, and the organization of two TempEval campaigns, has revived interest in temporal processing and triggered a shift to machine learning techniques. The standard tasks include the detection of events and timex, the anchoring of events to times, and the ordering of events restricted to selected contexts (sentences/consecutive sentences) [Verhagen *et al.*, 2010] or document-wide [Mani *et al.*, 2006]. Even when they address the task of event ordering at the document level, researchers simplify the problem to allow for the straightforward application of classification techniques. Thus, most research have focused on the task of independently predicting the correct relation for pre-selected pairs of events, explicitly in the human annotations. Going back to our example, this means only predicting for the unique E-E pair (e_5, e_{12}). It is furthermore usually assumed that the E-E pair is already ordered (i.e., (e_5, e_{12}) is a decision point, but (e_{12}, e_5) is not), thus reducing the number of possible relation labels from 13 to 6 [Mani *et al.*, 2006]. Temporal reasoning is sometimes invoked, but only during training and only to expand the pool of pair examples. [Mani *et al.*, 2006] use this resampling technique and report an accuracy score of 93.1% (62.5% without resampling) for the 6-way classification task. The majority class (i.e., *before*) baseline in this case is 75.2% (51.6% without resampling). An obvious shortcoming of this "local" approach is that it ignores the algebraic properties of temporal relations at prediction time, and does not guarantee the coherence of the event graph at the document level. We will see in Section 5 that most structures predicted this way are incoherent, therefore useless for downstream applications. More recently, various attempts have been made at predicting globally coherent structures. These approaches still rely on a locally-trained classifier but they use the algebraic properties of relations as constraints during inference. For example, [Tatu and Srikanth, 2008] propose a greedy search procedure with backtracking that is applied to a graph of events, in which the pairs are set to the reference pairs. In another context, [Bramsen *et al.*, 2006] describe various other greedy inference schemes. Finally, [Bramsen *et al.*, 2006; Chambers and Jurafsky, 2008] reformulate the problem of "decoding" a temporal graph under coherence constraints as a global optimization problem that can solved exactly with integer programming. Importantly, both papers use a small subset of the TimeML relations: only strict precedence relations (*before*, *after*). This has a clear advantage from a combinatorial perspective: by reducing the number of relations from 13 to 2, they manage to obtain ILP formulations that have a reasonable number of variables and constraints, at the expense of expressiveness. Chambers and Jurafsky [Chambers and Jurafsky, 2008]'s global model provides accuracy gains over the local classifier alone (from 66.8% to 70.4%), on the saturated gold event graph. But these improvements are partly obtained thanks to additional constraints derived from the gold (namely, all the E-T and T-T relations, which they saturated and supplemented with their automatically computed T-Ts). So it is unclear where these improvements come from between the global optimization and the E-T/T-T oracle. We will check this in our own experiments by directly measuring the peformance of this oracle. The approach proposed by [Bramsen *et al.*, 2006] is similar to [Chambers and Jurafsky, 2008], but its scope is different. They consider the task of ordering entire paragraphs (from the biomedical domain), so that the number of entities to relate is much smaller (they report 20 segments on average), but do not assume the reference pairs and are closer to the global task. Another related work is [Yoshikawa *et al.*, 2009], who also adopt a global approach that uses Markov Logic (instead of ILP) to jointly predict E-T and E-E relations (but only between entities appearing in consecutive sentences).

It should be clear that trying to generalize the approach of [Chambers and Jurafsky, 2008] to the full interval algebra is going to create an important combinatorial explosion both in terms of the number of variables and the number of constraints needed to represent the problem. In order to generalize to all TimeML relations, one has to consider every "generalized" relation and their combinations. The composition of any generalized relation, as we just saw, requires a specification of the 13×13 base compositions. The number of possible relations between two events is in general $2^{13} = 8192$, and 82 if one only considers convex relations. This translates into a minimum of $82 * n^2$ LP variables (for n entities), and $82^2 * n^3$ constraints (for relation compositions alone); that is, 8200 variables and close to 7 million constraints for just 10 entities. Section 3 comes back in detail on these aspects, but it is important to note that disjunctive information is crucial

for the checking of the consistency of predictions.

3 A global model over endpoints

Section 2 indicated that the global inference strategy used by [Bramsen *et al.*, 2006] and [Chambers and Jurafsky, 2008] is not likely to scale up when considering the full set of temporal relations in the TimeBank. This section proposes a simple solution to control this combinatorial complexity: to reformulate the optimization problem into a simpler yet semantically equivalent form by using the conversion between interval and interval endpoints.

3.1 Events and event endpoints

As for intervals, one can define an algebra on interval endpoints or Point Algebra (PA) [Vilain *et al.*, 1990]. By relying on a smaller relation set (and also fewer possible compositions), this algebra is better suited for inference purposes in our context. While there are 13 possible base relations on intervals, there are only 3 base relations (and only 7 when considering all the possible disjunctions). The basic relations are noted $\prec$, $\succ$, and =. The disjunctions are noted $\preceq$, $\succeq$, $\prec\succ$, and the general disjunction $\{\prec, =, \succ\}$, meaning no information is known ("vague"). We will use only 5 of these, since the vague relation is implied in absence of a more specific information, and $\prec\succ$ is impossible when we start with only basic (precise) relations on points. Translations from or into PA indeed preserve the convex property, and path-consistency is thus complete for the translation of a convex graph into PA. The resulting compositions are listed in Table 2; a blank indicates the result is the vague relation so this composition needs not be expressed. There are now only 17 constraints resulting from compositions for each event pair. Once the composition

∘	$\prec$	$\preceq$	$\succ$	$\succeq$	=
$\prec$	$\prec$	$\prec$			$\prec$
$\preceq$	$\prec$	$\preceq$			$\preceq$
$\succ$			$\succ$	$\succ$	$\succ$
$\succeq$			$\succ$	$\succeq$	$\succeq$
=	$\prec$	$\preceq$	$\succ$	$\succeq$	=

Allen	order/endpoints
b	$(\prec, \prec, \prec, \prec)$
m	$(\prec, =, \prec, \prec)$
o	$(\prec, \succ, \prec, \prec)$
s	$(=, \succ, \prec, \prec)$
d	$(\succ, \succ, \prec, \prec)$
f	$(\succ, \succ, \prec, =)$

Table 2: Composition of point relations (left) and correspondence events/points (right).

constraints are taken into account, we can translate back to a set of constraints on events. For each event pair (e_1, e_2), we consider the four relations between their endpoints , (e_1^-, e_2^-), (e_1^+, e_2^-), (e_1^-, e_2^+), (e_1^+, e_2^+) shown Table 2. If these relations are all specified and non disjunctive, it translates as a unique Allen base relation. If the point relations are disjunctive, the translation is the disjunction of all the translations obtained by distributing the point disjunctions: e.g. the four relations $(\preceq, \succ, \prec, \prec)$ correspond to the two possibilities $(\prec, \succ, \prec, \prec)$, $(=, \succ, \prec, \prec)$, translated as $overlaps$ or $starts$.

3.2 ILP formulation

We detail here the encoding of global constraints on events in Integer Linear Programming. ILP solves constraints while optimizing an objective function, a linear combination of a set of weighted variables. In our case, these variables correspond to triples (r, p, q), where p and q are endpoints, and r a possible relation over them. The weights on those variables, noted $c_{(r,p,q)}$, are derived from the classifier scores (on event pairs) by summing over the relevant interval relations on intervals containing those points. For instance, the score for $\{\prec, p, q\}$ is obtained by summing the scores for relations before, meet, duringi, and finishi for the event pairs that includes p and q as endpoints.

Formally, let P be the set of points resulting from the translation of events. Let $R = \{\prec, \succ, =\}$, and $R^+ = R \cup \{\preceq, \succeq\}$. Let also Inv: $R \rightarrow R$ be the inverse operator, and Comp: $(R \times R) \rightarrow R$ the composition operator.

The objective function is defined as:

$$max \sum_{(p,q)\in C} \sum_{r\in R} c_{(r,p,q)} \cdot x_{(r,p,q)} \quad (1)$$

$$x_{(r,p,q)} \in \{0,1\} \quad \forall \langle p,q\rangle \in P \times P, p \neq q, \forall r \in R^+ \quad (2)$$

Note that for n initial events, there are $2n$ points, so we end up with $5 * (2n)^2 = 20n^2$ variables (i.e., 4 times less than for the equivalent event-based formulation).

We now put the following additional constraints on these integer variables. For all pairs of points e_i^-, e_i^+ defining the event e_i , we set $x_{(\prec,e_i^-,e_i^+)} = 1$ and $x_{(=,e_i^-,e_i^+)} = x_{(\succ,e_i^-,e_i^+)} = 0$.

At most one base relation can hold between 2 points:

$$\sum_{r\in R} x_{(r,p,q)} \leq 1 \qquad \forall \langle p,q\rangle \in P \times P, p \neq q \quad (3)$$

Disjunctive relations are related to simple relations:

$$x_{(\prec,p,q)} + x_{(=,p,q)} \leq x_{(\preceq,p,q)} \quad (4)$$

$$x_{(\succ,p,q)} + x_{(=,p,q)} \leq x_{(\succeq,p,q)} \quad (5)$$

In case compatible relations $x_{(\preceq,p,q)}$ and $x_{(\succeq,p,q)}$ are separately inferred, we must ensure the most specific information and incompatible relations cannot be inferred at the same time, which is achieved with:

$$x_{(\preceq,p,q)} + x_{(\succeq,p,q)} \leq x_{(=,p,q)} + 1 \quad (6)$$

$$x_{(\preceq,p,q)} + x_{(\succ,p,q)} \leq 1 \quad (7)$$

$$x_{(\prec,p,q)} + x_{(\succeq,p,q)} \leq 1 \quad (8)$$

These correspond to the "intersections" of relations in the algebra. Also, we add symmetry constraints:

$$x_{(r,p,q)} \leq x_{(Inv(r),q,p)} \qquad \forall r \in R+ \quad (9)$$

$$x_{(r,p,q)} \geq x_{(Inv(r),q,p)} \qquad \forall r \in R+ \quad (10)$$

Finally, the triangular constraints from Table 2 are, $\forall \langle p,q,t\rangle \in P \times P \times P, p \neq q \neq t, \forall r_1, r_2 \in R^+$:

$$x_{(r_1,p,q)} + x_{(r_2,q,t)} \leq x_{(Comp(r_1,r_2),p,t)} + 1 \quad (11)$$

These represent the bulk of our constraints: they are in the order of $|P|^3 \times |R^+|^2$, ie $8 * 17n^3$ if n is the number of events ($|P| = 2n$). The equivalent generalized (convex) event based version has $82 * 82n^3$ constraints, about 50 times more, and would need a huge additional set of intersective constraints per edge (*vs.* 3 in PA).

4 Graph decomposition

Considering all possible relations on the graph of events puts a considerable burden on the decoding phase, and the translation into PA partially addresses this problem. But considering all possible E-E pairs has also a strong (potentially negative) impact on the precision of the predictions. As it stands, our ILP formulation is biased toward predicting simple relations. This is due to the fact that disjunctive relations are not part of the objective function (the classifier does not assign them any score), so disjunctions are only going to be predicted in cases where none of the base relations produce a coherent structure. One possible way within ILP to avoid this "overzealous" behavior of the system is to add a "vague" class to all unrelated event pairs and add them to the training phase, as in [Chambers and Jurafsky, 2008], but this leads to an overrepresentation of the vague relations. Another solution is to find a decomposition of the temporal graph so as to limit the decisions to relevant subgroups of events, and predict relations only *within* these subgroups. This assumes it is possible to structure temporal situations into so-called time-frames that help mental representations. This is in a way the mirror image of the idea proposed in [Bramsen *et al.*, 2006], where consecutive events are grouped within large segments, and relations are predicted *between* these segments. The decomposition must be meaningful for that strategy to succeed, and we must find a way of adding between-groups relation afterwards. We tried two different decompositions. The first is based on an observation: human annotations are often scattered, and events appear in separate, smaller, self-connected components. Using the structural knowledge encoded in the gold as an oracle, we restrict our constrained prediction strategy to these connected components. No other knowledge from the gold was considered (in particular, no knowledge as to which events were related in the gold).

The second decomposition groups events with dates appearing in the same sentence, exploiting the fact that most dates in TimeBank have determined values. Events located in sentences without dates were arbitrarily attached to the most recently introduced date. We then predict relations within each subgroup with a consistent prediction; finding relations between subgroups boils down to finding relations between the dates they are centered on, and that information is already available from the extraction phase we assumed, as in the previous experiments.

5 Experiments

We carried out two main sets of experiments to evaluate our global ILP model on endpoints. First, we evaluated the bare model presented in Section 3 on the fully closed reference event graph, i.e. the graph resulting from saturating all E-E, E-T, and T-T annotations (including the extra relations from Bethard, and those obtained by date calculations). Secondly, we combine the ILP point model with the graph decompositions described above. All our experiments were performed using 5-fold cross-validation on the TimeBank 1.2 corpus.

The ILP-based system (ILP) is compared with a number of other systems: the base classifier without consistency checking, a baseline ordering events in the order of the text (BEFORE), a greedy natural reading ordering taking the most probable relations in sequence (NRO), and an oracle (ZERO-EE) that provides E-E relations predicted only by saturation over the E-T and T-T relations we start with. As our base model, we used a log-linear (aka Maxent) 11-way classifier trained on the fully closed TimeBank (each E-E pair of the graph was used as training instance). The feature set relies on the attributes provided in the TimeBank (event class, tense, aspect) as in [Mani *et al.*, 2006; Chambers and Jurafsky, 2008]. Parameters were estimated using the Megam package[2] under the default settings. For training, the data was sampled from the TimeBank in a way that mirrors the decompositions; no resampling was performed for the first experiment. The LP solver we used is SCIP[3], which is currently the fastest non-commercial mixed integer programming solver. The solver was timed out after 1 hour or stopped when it reached 2×10^6 triangular constraints. These cases are reported as inconsistent output in the results. The evaluation is always made w.r.t. to all *simple* relations inferrable from the reference annotations. This only makes sense if the resulting reference graph is consistent, and we thus restricted this evaluation to the 139 texts where this is true (out of 186). We use micro-average to balance the contributions made by short and long texts.

Results for the first experiment are shown in Table 3. We report scores in terms of accuracy, since we assume the reference event graph, to match experiments in [Chambers and Jurafsky, 2008]. The systems also generate different inferences and produce simple relations on pairs not in the reference, and this will be evaluated with recall/precision in the context of the more complete second experiment.

Making zero E-E predictions and assuming only E-T relations (before saturation of the graph) yields an accuracy of 26%; this is to be considered as the minimum information provided before predicting E-E relations. The relatively high accuracy of this oracle method should however be taken with caution, as it assumes perfect knowledge of E-T relations, while the best dedicated systems in simpler settings reach accuracies of 80-90% [Verhagen *et al.*, 2010; Mani *et al.*, 2006]. We tested that by changing at random 10% of these assumed relations, a lot of inconsistencies arise and accuracy drops to 5%. Accuracy here is actually a recall measure: [Chambers and Jurafsky, 2008] don't consider predictions made by their system (even if it's a precise precedence relation) when the reference does not imply a *before/after* relation. The importance of this can be seen from the fact that a simple consistent baseline as BEFORE reaches about 38% of recall while generating 94% of erroneous predictions. The consistent NRO fares better on precision (60%), at the expense of a drop in accuracy on the reference. The base classifier has an accuracy on separate decisions of 60.1% (trained and tested on saturated data), but 82% of the graph it predicts are inconsistent. When ILP is used to constrain these predictions we reach a recall of almost 50%. This is the score reached by [Tatu and Srikanth, 2008] on the reference predictions where they actually managed to preserve consistency, and we use the whole set of relations while they have only six.

[2] http://www.cs.utah.edu/~hal/megam/

[3] http://scip.zib.de/

In a comparable setting but with only two relations, [Chambers and Jurafsky, 2008] reach 70% accuracy.

System	ZERO-EE	BEFORE	NRO	ILP
Accuracy	26.01	37.93	20.08	49.80

Table 3: Results on TB, assuming reference event pairs

We now turn to the much harder task of predicting the temporal structure without pre-selecting event pairs; performance scores with a decomposition in self-connected maximal components are given in Table 4. We repeat BEFORE scores for clarity, although it is by construction insensitive to the different setting (as ZERO-EE).
On the connected component oracle, we can see that NRO holds its ground even without selecting event pairs, and ILP benefits the most of the context, even though it times out on a few texts. We estimated that ILP was practical up to 25 nodes in a sub-graph. The base classifier predicts 88% inconsistent graphs here. Results on decomposition around dates are of

System	Precision	Recall	F1-score	Inco.
ILP	33.02	54.07	41.00	5.93
NRO	49.98	17.02	25.40	0.00
BEFORE	6.22	37.93	10.69	0.00

Table 4: Decomposition on connected components

course lower since no knowledge of E-E relations is provided. NRO and ILP are close (respectively 16.3% and 15.0% in F1), and again ILP has a better recall while NRO has a better precision. The percentage of inconsistent predictions by ILP is now up to 8.6%, due to the absence of consistency constraints between subgroups. This is less than 3% more than the number of time-outs, an indication that this decomposition is not too damaging for consistency. Assuming this decomposition the base classifier produces now 96% of inconsistent graphs. Our decomposition method is clearly too crude, and can only serve as a starting point for more elaborate approaches. We can nonetheless claim that the problem of temporal structure prediction is doable within this framework, without assuming anything about E-E relations, or resampling the test environment. For each experiment, we tested the differences between ILP and the other methods with a Wilcoxon signed-rank test over the measures for each text and found high significance levels ($p < 10^{-5}$ at worst).

6 Conclusion

We have generalized proven strategies for temporal prediction from a few simple relations to the complete set of relations used in existing temporally annotated corpora like TimeBank. We also provide the definition of intermediate steps in the direction of predicting temporal structures without assuming too much of the target representation, mainly without assuming the small subset of event pairs to relate among the n^2 possible pairs. This is done by looking at decompositions of the temporal structure, either using some knowledge about the gold annotation, or using a basic heuristics, and to the best of our knowledge the latter is the first attempt at the unrestricted global task on a significative corpus of texts. It has room for improvement, since we did not use any linguistic knowledge about global attachment of temporal entities. Event grouping could be a relevant subtask for temporal extraction.
The use of point based representations could be pushed further in future work to check whether it can also help the learning phase, by translating the corpus before training. We have not devoted too much attention to the learning phase here, but in the context of a global prediction, we could test on the utility of learning to predict not only simple relations (since this forces a lot of decisions to be made) but also vaguer ones (in a manner similar to the final experiment of [Chambers and Jurafsky, 2008]).

References

[Allen, 1983] J. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 1983.

[Bethard *et al.*, 2007] S. Bethard, J. H. Martin, and S. Klingenstein. Timelines from text: Identification of syntactic temporal relations. In *Proc. of ICSC*, 2007.

[Bramsen *et al.*, 2006] P. Bramsen, P. Deshpande, Y. K. Lee, and R. Barzilay. Inducing temporal graphs. In *Proc. of EMNLP*, 2006.

[Chambers and Jurafsky, 2008] N. Chambers and D. Jurafsky. Jointly combining implicit constraints improves temporal ordering. In *Proc. of EMNLP*, 2008.

[Mani *et al.*, 2006] I. Mani, M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovsky. Machine learning of temporal relations. In *Proc. of ACL*, 2006.

[Pustejovsky *et al.*, 2005] J. Pustejovsky, R. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. The specification language TimeML. In I. Mani, J. Pustejovsky, and R Gaizauskas, editors, *The Language of Time: A Reader*. OUP, 2005.

[Tatu and Srikanth, 2008] M. Tatu and M. Srikanth. Experiments with reasoning for temporal relations between events. In *Proc. of Coling*, 2008.

[Van Beek, 1990] P. Van Beek. *Exact and approximate reasoning about qualitative temporal relations*. PhD thesis, University of Waterloo, Canada, 1990.

[Verhagen *et al.*, 2010] M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proc. of the 5th International Workshop on Semantic Evaluation*, 2010.

[Verhagen, 2005] M. Verhagen. Temporal closure in an annotation environment. *Language Resources and Evaluation*, 39(2), 2005.

[Vilain *et al.*, 1990] M. Vilain, H. Kautz, and P. van Beek. Constraint propagation algorithms for temporal reasoning: a revised report. 1990.

[Yoshikawa *et al.*, 2009] K. Yoshikawa, S. Riedel, M. Asahara, and Y. Matsumoto. Jointly identifying temporal relations with markov logic. In *Proc. of ACL*, 2009.